新模糊聚类有效性指标 *

耿嘉艺，钱雪忠，周世兵

(江南大学 物联网工程学院，江苏 无锡 214122)

摘　要：模糊聚类是模式识别、机器学习和图像处理等领域的重要研究内容。模糊 C-均值聚类算法是最常用的模糊聚类实现算法，该算法需要预先给定聚类数才能对数据集进行聚类。提出了一种新的聚类有效性指标，对聚类结果进行有效性验证。该指标从划分熵、隶属度、几何结构角度，定义了紧凑度、分离度、重叠度三个重要特征测量。在此基础上，提出了一种最佳聚类数确定方法。将新聚类有效性指标和传统有效性指标在 6 个人工数据集和 3 个真实数据集进行实验验证。实验结果表明，所提出的指标和方法能够有效地对聚类结果进行评估，适合确定样本的最佳聚类数。

关键词：模糊 C-均值聚类；聚类数；聚类有效性指标；模糊聚类

中图分类号：TP391

New fuzzy clustering validity index

Geng Jiayi, Qian Xuezhong, Zhou Shibing

(*Shool of Internet of Things Engineering*, *Jiangnan University*, *Wuxi Jiangsu* 214122)

Abstract: Fuzzy clustering is an important research content in the fields of pattern recognition, machine learning and image processing. Fuzzy C-means clustering algorithm is the most commonly used fuzzy clustering algorithm. The algorithm needs to preset the number of clusters in order to cluster the data set. This paper propose a new clustering validity index to validate the clustering results. This index defines the three important features of compactness, resolution and overlap degree from the perspective of partition entropy, membership degree and geometric structure. On this basis, this paper propose a method of determining the optimal clustering number. This paper validate the new clustering validity index and the traditional effectiveness index in six artificial data sets and three real data sets. The experimental results show that the proposed indexes and methods can effectively evaluate the clustering results and are suitable for determining the optimal clustering number of the samples.

Key Words: fuzzy C-means clustering; number of clusters; clustering validity index; fuzzy clustering

0　引言

聚类是将没有先验知识的样本，按照特定的规则，将相似的样本归为一类，不相似的样本分到不同的类中[1]。聚类分为两大方向，传统聚类和模糊聚类。传统聚类为硬划分，每个样本必须清晰的划分到不同的子类中，只有属于和不属于两种情况。但是现实中的大部分数据都具有不确定性，一个样本数据可能在不同程度上属于多个类[2]。因此，Ruspini[3]引入了模糊划分的概念，从而出现了模糊聚类。相应地隶属度范围也从二值逻辑{0,1}扩展到[0,1]。模糊聚类相比传统聚类，更能反映出真实的世界。实现模糊聚类最常用的算法为 Dunn[4]提出的模糊 C-均值算法（fuzzy C-Means，FCM）。该算法通过迭代，使目标函数最小化。FCM 算法设计简单、解决问题的范围广。但是 FCM 算法需要通过聚类有效性验证，以确定最佳聚类数和判断分类结果的好坏[5]。

选择合适的聚类有效性指标是研究聚类有效性的重要步骤[6]。目前已经存在了许多聚类有效性指标，但是由于数据集的结构多种多样，没有一个聚类指标适用于任何类型的数据集，没有一种指标的表现总优于其他指标[7]。比如朴尚哲等在文献[8]中,列举了近年来一些常用的聚类有效性指标，包含基于隶属度的聚类有效性指标、基于类内紧致度和类间离散度的聚类有效性指标、基于熵和数据结构的聚类有效性指标等，这些指标只能在特定的数据集上发挥自己的优势，并不能运用在所有数据集上。本文针对现有模糊聚类有效性指标的不足，提出新的聚类有效性指标。该指标结合数据集的划分熵、隶属度以及数据结构，定义了紧凑度、分离度、重叠度，能够克服噪声和重叠的影响，准确地找到最佳聚类数。实验结果表明，新指标在人工数据集和真实数据集上均取得了良好的效果。

基金项目：国家自然科学基金资助项目（61673193）；中央高校基本科研业务费专项资金资助项目（JUSRP11235，JUSRP51635B）

作者简介：耿嘉艺（1992-），女，山西晋中人，硕士研究生，主要研究方向为模式识别、机器学习研究；钱雪忠（1967-），男，江苏无锡人，副教授，硕导，主要研究方向为数据挖掘、机器学习；周世兵（1972-），男，江苏盐城人，讲师，博士，主要研究方向为模式识别、人工智能.

1　相关工作

1.1　FCM 算法

FCM 算法是基于目标函数的模糊 C 划分，通过优化目标函数得到均匀的 c 个模糊集[9]。目标函数是由隶属度、样本到聚类中心的偏差，两者结合构成。通过迭代，最小化目标函数，当迭代次数超过规定的数值或目标函数差值小于阈值时终止。FCM 需要事先初始化聚类原型和给定聚类数。

假设数据集有 n 个样本，每个样本为 p 维 $X=\{x_1,x_2,...,x_n\},X\in R^p$。目标函数为

$$J_m(U,V,X)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\|x_k-v_i\|^2 \tag{1}$$

隶属度矩阵约束条件：

$$0\leqslant u_{ij}\leqslant 1,1\leqslant i\leqslant c,1\leqslant j\leqslant n$$

$$\sum_{i=1}^{c}u_{ij}=1,1\leqslant j\leqslant n$$

$$0\leqslant\sum_{j=1}^{n}u_{ij}\leqslant n,1\leqslant i\leqslant c$$

其中：c 表示聚类数；m 表示模糊程度，范围[1,∞]；U 为 c*N 矩阵，表示样本属于模糊子集的隶属程度；V 为 c*p 矩阵，表示聚类原型。FCM 算法通过不断迭代更新聚类原型 V 和 U，从而最小化目标函数。更新公式为：

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^{m}x_k}{\sum_{k=1}^{n}u_{ik}^{m}},1\leqslant i\leqslant c \tag{2}$$

$$u_{ik}=\frac{\|x_k-v_i\|^{-\frac{2}{m-1}}}{\sum_{i=1}^{c}\|x_k-v_i\|^{-\frac{2}{m-1}}}$$,1�i�c,1�k�n　(3)

FCM 算法步骤如下：

a)给定参数 C、模糊程度 m、最大迭代次数和迭代终止条件。

b)初始化聚类原型，并更新模糊隶属矩阵 U。

c)更新模糊聚类原型矩阵 V。

d)如果大于迭代次数或目标函数差值小于阈值，则停止，否则转到步骤 b)。

1.2　传统聚类有效性指标

1) PC[10]

$$V_{PC}=\frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{2} \tag{4}$$

划分系数 PC 形式简单，易于计算，但是仅考虑了每个集群的紧凑度，并且与数据的几何结构缺乏直接的联系。随着聚类数的变化，呈现单调趋势。这些不足直接导致指标无法验证具有大量小簇的分区和复杂数据集。

2) MPC[11]

$$V_{MPC}=1-\frac{k}{k-1}(1-V_{PC}) \tag{5}$$

指标对划分系数PC存在的单调递减趋势问题进行了优化，但是对于 PC 指标其他方面地缺陷并没有进行改进。指标在人工数据集上的效果不理想。

3) PE[12]

$$V_{PE}=-\frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}\log u_{ij} \tag{6}$$

划分熵 PE 指标简单，运算量小。同样存在以下问题：只考虑了每个集群的紧凑度；与数据集的几何结构缺乏联系；存在单调趋势。指标仅在分离较好的数据集上，表现良好，在噪声和重叠数据集上表现不佳。

4) XB[13]

$$V_{XB}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|v_i-x_j\|^2}{n\min_{i\neq j}\|v_i-v_j\|^2} \tag{7}$$

指标将数据的隶属度和几何结构考虑在内，紧致度为所有样本数据到聚类中心距离的和，分离度为类中心之间距离的最小值。该指标存在两个缺点：当 c→n 时，XB 指标变为 0；当 m→∞时，XB→∞。在上述两种情况下，指标失去稳定性，无法判断最佳聚类数。

5) UV[14]

$$V_{UV}=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{2}+\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{2}\exp(-\frac{\|x_j-v_i\|^2}{\varepsilon}) \tag{8}$$

$$\varepsilon=\frac{1}{n}\sum_{j=1}^{n}\|x_j-\bar{x}\|^2,\bar{x}=\frac{1}{n}\sum_{j=1}^{n}x_j$$

该指标引入指数函数测量数据与中心的距离，相比于欧氏距离，在一定程度上能够克服噪声对数据的影响，但是由于指标只考虑了集群的紧凑度和分离度，没有考虑重叠度对分类的重大影响，所以在重叠数据集上效果不是很理想。

6) FM[15]

$$V_{FM}=\alpha_f\times(-\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}[u_{ij}\log(u_{ij})]) \tag{9}$$

$$\alpha_f=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij}-1/c)^2\|x_j-v_i\|^2}{n\min_{i\neq k}\|v_i-v_k\|^2}$$

该指标将划分熵和模糊划分因子这两个重要评价指标考虑在内，定义了聚类的紧致性和分离性。由于该指标采用相距最近两个类中心之间的距离作为的分离度，此类情况在噪声数据集上的表现不佳。

2　新聚类有效性指标

聚类有效性指标的好坏，直接影响着最终聚类结果的质量。新聚类有效性指标是由紧凑度、重叠度和分离度三者共同构成。紧凑度由类内距离表示，分离度由最小隶属度表示，重叠度由隶属度和划分熵相结合表示。好的聚类对应较小的紧凑度、重

叠度，较大的分离度。该指标充分考虑了数据集的整体信息，能够准确地判断出数据集的最佳聚类数。

2.1 紧凑度

定义 1 定义 vs(c,U)为第 i 类所有样本到第 i 类中心的平均距离，即

$$vs(c,U)=\sum_{i=1}^{c}\sum_{k=1}^{n(i)}\frac{\|x_k-v_i\|^2}{n(i)} \tag{10}$$

其中：x_k 表示第 i 类的第 k 个样本，v_i 表示第 i 类的聚类中心，$n(i)$ 表示第 i 个聚类的样本数目。

定义 2 定义 vd(c,U)为第 i 类所有样本两两之间的平均距离，即

$$vd(c,U)=\sum_{i=1}^{c}\sum_{k\neq h}^{n(i)}\frac{\|x_k-x_h\|^2}{(n(i)^2-n(i))/2} \tag{11}$$

其中：分母表示第 i 类所有样本两两之间距离的总个数。

定义 3 定义类内紧凑度 Var(c,U)为前两者相加（样本与中心距离的平均值、类内样本之间距离的平均值）的和，即：

$$Var(c,U)=vs(c,U)*vd(c,U)$$

$$=\sum_{i=1}^{c}\sum_{k=1}^{n(i)}\frac{\|x_k-v_i\|^2}{n(i)}*\sum_{i=1}^{c}\sum_{k\neq h}^{n(i)}\frac{\|x_k-x_h\|^2}{(n(i)^2-n(i))/2} \tag{12}$$

紧凑度表示类内样本的集中程度。为了解释相关概念，结合示意图进行说明。图 1 表示第 i 类的所有样本点到该类聚类中心的距离，值越小，说明类内样本距离类中心越近，表明了类内样本与类中心的结构关系；图 2 表示的第 i 类所有样本数据，两两之间的距离，值越小，说明类中的数据越紧，表明了类内样本数据的整体结构信息。类内紧凑度将两者结合起来，共同发挥各自的优势。显然 Var(c,U)的最小值，表明类内的数据点彼此接近，具有较高的紧凑度。

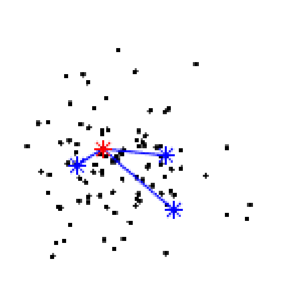


图 1 类内样本与中心距离　　图 2 类内样本之间距离

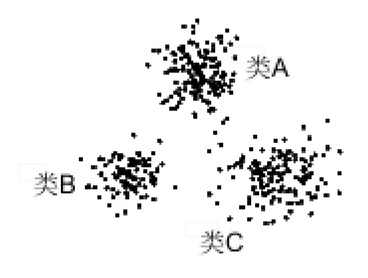


图 3 三个类分布示意图

2.2 分离度

定义 4 定义 S_{ij} 为第 k 个样本属于第 i 类和第 j 类之间最小的隶属度，即：

$$S_{ij}=\min_{i\neq j}(u_{ik},u_{jk}),k=1,2,...,n \tag{13}$$

定义 5 定义总体离散度 Sep(c,U)为 S_{ij} 总和的相反数，即

$$Sep(c,U)=1-\frac{1}{c}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}(\frac{1}{n}\sum_{k=1}^{n}S_{ij}) \tag{14}$$

分离度表示两个模糊集群之间的分离程度。大部分指标的分离度通过计算类中心之间的距离，而类中心无法反映样本分布的整体形状，而且对于噪声数据使用距离判断会出现偏差。比如在图 3 中，两个类之间具有相同的距离，分离性也可以不同。AB 与 AC 类中心的距离相等，但是明显 AB 比 AC 分离。此处新指标借鉴 Chen 等[16]提出的分离度，通过一个样本数据相对两个类的模糊隶属度最小值作为分离度量，第 k 个样本越接近于一个类的距离中心，则相对该类的隶属度越接近于 1，另一个类越接近于 0，相应地定义 4 的值也越接近于 0。此时类间模糊性也越小，类间越分离。总体离散度是将定义 4 的值求和，并取反求得，值越大，表示数据点相对于类的模糊程度越低，越能够清晰地划分到集群中，分离越好。

2.3 重叠度

定义 6 定义 C_{ij} 为第 k 个样本属于第 i 类和第 j 类之间隶属度的乘积，即

$$C_{ij}=\sum_{k=1}^{n}(u_{ik}^2*u_{jk}^2),k=1,2,...,n \tag{15}$$

定义 7 定义总体重叠度 Cop(c,U)为 C_{ij} 的总和与熵的结合,即

$$Cop(c,U)=\frac{1}{n}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}C_{ij}*f(x_k) \tag{16}$$

$$f(x_k)=-\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}*\log u_{ik}$$

重叠度用于衡量界限不明确的两个类之间的重叠程度。两个类之间的重叠度定义为隶属度平方的乘积，当两个类别之间划分较清晰，隶属度之间相差越大，乘积的值越小，类划分时越明确，聚类结果越清晰。第 k 个样本相对于每类样本的隶属度都为 1/c，此时重叠度的值达到最大。这里与熵结合起来，可较好地反映出划分结果的模糊程度和不确定性度，值越小，它的不确性程度越小，需要的信息量越小，则分类效果越可靠，此处作为权重评价指标。显然总体重叠度的值越小，说明两个类之间划分越清晰，重叠越小。

2.4 归一化

由于紧致度、分离度和重叠度有不同的量纲，故需做归一

化处理，将各个聚类数对应地的指标值，除以最大的指标值，此时各度量范围变为[0,1]，其结果可表示为

$$Var^{n}(c,U)=\frac{Var(c,U)}{Var_{max}},Var_{max}=\max_{c}(Var(c,U)) \quad (17)$$

$$Sep^{n}(c,U)=\frac{Sep(c,U)}{Sep_{max}},Sep_{max}=\max_{c}(Sep(c,U)) \quad (18)$$

$$Cop^{n}(c,U)=\frac{Cop(c,U)}{Cop_{max}},Cop_{max}=\max_{c}(Cop(c,U)) \quad (19)$$

2.5 聚类有效性指标

$$W(c,U)=Var^{n}(c,U)+\frac{Cop^{n}(c,U)}{Sep^{n}(c,U)} \quad (20)$$

将划分熵、隶属度、几何结构这些模糊聚类中重要的特征结合起来，共同构成新聚类有效性指标。从紧凑度角度出发，希望 $Var(c,U)^{\{n\}}$越小越好，表示类内距离越紧密；从分离度出发，希望 $Sep(c,U)^{\{n\}}$越大越好，表示类间距离越分散；从重叠角度出发，希望 $Cop(c,U)^{\{n\}}$越小越好，表示重叠达到最小。显然，W(c,U)越小，表示数据点被清晰地分到集群中，此时聚类效果最好。最佳聚类数与数据集的真实结构相符，找到最佳聚类数是聚类有效性指标的首要任务。该指标在噪声数据集、重叠数据集上，都能够准确地找到最佳聚类数。

3　确定最佳聚类数的算法

本文是在 FCM 算法和 W 聚类有效性指标下，提出的一种新的确定最佳聚类数的算法，解决了 FCM 需要事先确定最佳聚类数的问题，步骤如下：

a)初始化聚类数 c 的选择范围为[Cmin、 Cmax]。

b)c 以 1 为单位递增，调用 FCM 算法，利用 FCM 得到的最优解（U,V）,带入 W 指标。

c)计算并存储聚类有效性指标的值。

d)如果 c<Cmax,c=c+1，转到步骤 2，否则转到步骤 5。

e)选取与最小指标值对应地 c 作为最佳聚类数。

f)输出最佳聚类数以及指标值。

4　实验结果

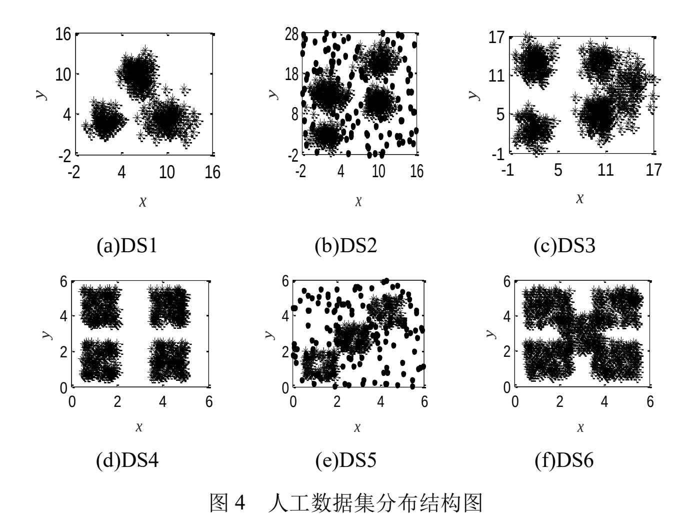
为了检验新聚类有效性指标能否取得良好效果，将新聚类有效性指标和已有的聚类有效性指标 PC、MPC、PE、XB、UV、FM，应用于 6 个人工数据集和 3 个真实数据集，观察它们的聚类效果。聚类数搜索范围为[2,Cmax]，Cmax=$\sqrt{n}$。指标中涉及的距离度量均为欧氏距离；参数 m 在何值取得最佳，尚缺乏理论指导，Pal 和 Bezdek[17]提出 m 在[1.5，2.5]时 FCM 聚类算法的结果最好，实验首先取 m=2 的情况。并且在不同的模糊加权 m 值下，观察新聚类有效性指标是否鲁棒。

4.1　有效性实验

4.1.1 人工数据集

图 4 为人工数据集分布结构示意图。DS1、DS2、DS3 是高斯分布数据集，DS4、DS5、DS6 是均匀分布数据集。DS1 和 DS4 数据集类与类之间分离明确；DS2 和 DS5 数据集有 150 个噪声污染数据；DS3 数据集中，三类样本数据彼此之间存在较大的重叠，另两类分离较好；DS6 数据集中，五类样本数据彼此之间都存在一定的重叠。

表 1 为人工数据集的具体数值信息和各个有效性指标计算得到的最佳聚类数，为了更直观地说明，图 5 详细地列出了 DS2 和 DS3 数据集的聚类数-指标关系图。针对分离较好的数据集 DS1 和 DS4，所有指标均有效。针对噪声数据集：DS2 数据集，PE 和 FM 得到的最佳聚类数为 2 类，其次才是 4，XB 则误判为 6 类，其余指标均可以有效的判定为 4 类；DS5 数据集，仅有 MPC、W 能正确的判定为 3 类。针对重叠数据集：DS3 和 DS6 由于重叠区域较多，传统聚类有效性指标失去判别能力，仅有 W 指标能够正确地判断这两个数据集的最佳聚类数为 5 类。



(a)DS1　(b)DS2　(c)DS3

(d)DS4　(e)DS5　(f)DS6

图 4　人工数据集分布结构图

表 1　7 种有效性指标在人工数据集上的最佳聚类数（m=2）

数据	样本数目	样本维数	实际类数	最佳聚类数 PC	MPC	PE	XB	UV	FM	W
DS1	403	2	3	**3**	**3**	**3**	**3**	**3**	**3**	**3**
DS2	653	2	4	**4**	**4**	2	6	**4**	2	**4**
DS3	603	2	5	4	4	2	4	4	2	**5**
DS4	600	2	4	**4**	**4**	**4**	**4**	**4**	**4**	**4**
DS5	390	2	3	2	**3**	2	2	2	2	**3**
DS6	750	2	5	2	4	2	4	4	2	**5**

4.1.2 真实数据集

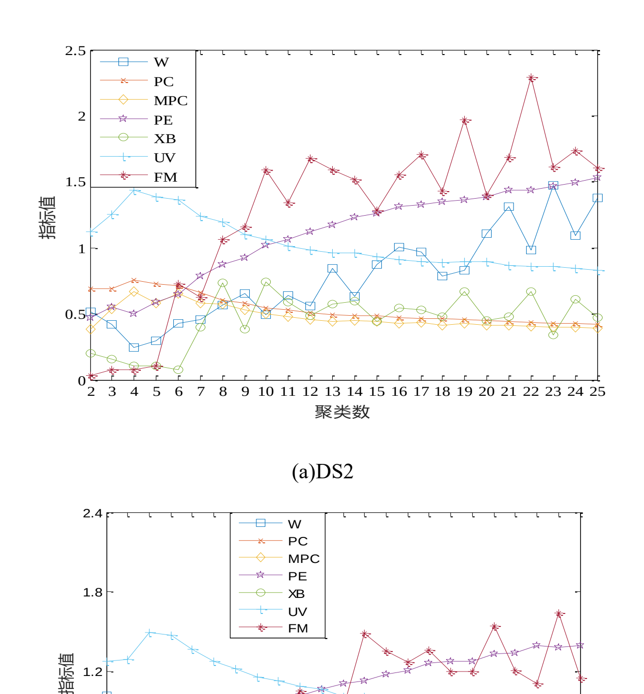
真实数据集来源于公共数据库 UCI 数据库，是加州大学欧文分校提出的，用于机器学习常用的标准测试数据库。

a)Iris 数据集：分为 3 类，分别为 Iris Setosa、Iris Versicolour、Iris Virginica。在这个数据集上，有两类数据几乎不可辨别，另一个集群分离较好。所以,最佳聚类数判定为 3 类，次最佳为 2 类，这两种情况符合数据集的结构。传统指标判定 2 类为最佳，新聚类指标 3 类为最佳。

b)Wdbc 数据集：分为 2 类，分别为 Malignant、Benign。

特征是从乳房块的细针抽吸（FNA）的数字化图像计算的，描述了图像中存在的细胞核的特征。核特征提取用于乳腺肿瘤诊断。以上所有指标都正确的判断出最佳聚类数为 2。

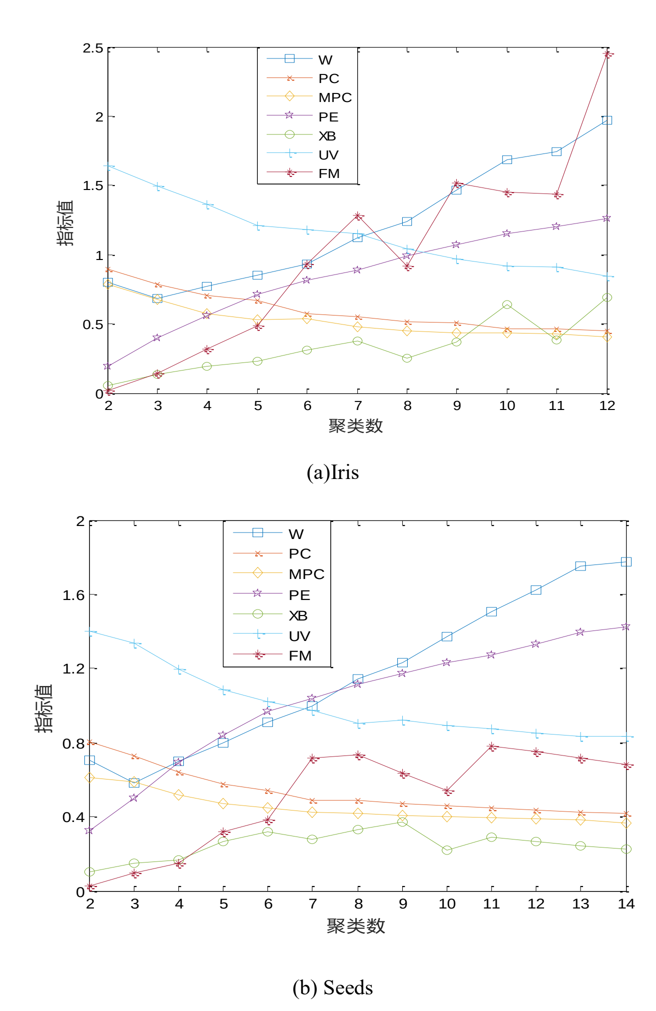
c)Seeds 数据集：分为 3 类，分别为三种不同品种的小麦籽粒 Kama、Rosa 和 Canadian。只有新聚类有效性指标能正确地判断出聚类数为 3 类，其余指标均误判为 2 类最佳。



(a)DS2

(b)DS3

图 5　7 种有效性指标的聚类数-指标关系图



(a)Iris

(b) Seeds

图 6　7 种有效性指标的聚类数-指标关系图

表 2 为真实数据集的具体数值信息和各个有效性指标计算得到的最佳聚类数，为了更直观地说明，图 6 详细地列出了 Iris 和 Seeds 数据集的聚类数-指标关系图。实验结果表明只有新聚类有效性指标可以在以上所有人工数据集和真实数据集下，正确判断出最佳聚类，在噪声和重叠数据都表现出了良好的效果。PC、PE、MPC 指标缺乏跟数据结构的直接联系，因而得到的最佳聚类有效性指标对应的聚类个数与实际情况不符。XB 指标当 m 和 c 增加到一定程度，失去可靠性。UV 和 FM 只考虑了紧凑度和分离度，没有考虑重叠度。新指标在一定程度上弥补了以上传统指标的缺点，具有较强的适应性。

表 2　7 种有效性指标在真实数据集上的最佳聚类数（m=2）

数据	样本数目	样本维数	实际类数	最佳聚类数 PC	MPC	PE	XB	UV	FM	W
Iris	150	4	3	2	2	2	2	2	2	**3**
Wdbc	569	32	2	**2**	**2**	**2**	**2**	**2**	**2**	**2**
Seeds	210	7	3	2	2	2	2	2	2	**3**

4.2　m 的比较

m 为模糊加权指数，控制着聚类结果的模糊程度，正是由于 m 的引入，使传统聚类推广到模糊聚类。m 根据已有经验一般选择范围在[1.5，2.5]。现将聚类有效性指标分别在 m= 1.5、1.7、2、2.3、2.5 五种情况下，应用于以上数据集。实验结果表明，PC、MPC、PE、XB 和 UV，随着 m 的变化，聚类结果发生变化，只有 FM 和 W 指标不随 m 的变化而变化，对 m 鲁棒；在聚类正确率方面，只有新提出的 W 指标，正确率为 100%。实验表明新指标在不同的 m 下，能够得到较好结果，有较强的可靠性和鲁棒性。

表 3　W 在不同模糊加权指数下的最佳聚类数

m	DS1	DS2	DS3	DS4	DS5	DS6	Iris	Wdbc	Seeds
1.5	3	4	5	4	3	5	3	2	3
1.7	3	4	5	4	3	5	3	2	3
2	3	4	5	4	3	5	3	2	3
2.3	3	4	5	4	3	5	3	2	3
2.5	3	4	5	4	3	5	3	2	3

表 4　PC 在不同模糊加权指数下的最佳聚类数

m	DS1	DS2	DS3	DS4	DS5	DS6	Iris	Wdbc	Seeds
1.5	3	4	4	4	2	4	2	2	2
1.7	3	4	4	4	2	4	2	2	2
2	3	4	4	4	2	2	2	2	2
2.3	3	4	2	4	2	2	2	2	2
2.5	3	4	2	4	2	2	2	2	2

表 5　MPC 在不同模糊加权指数下的最佳聚类数

m	DS1	DS2	DS3	DS4	DS5	DS6	Iris	Wdbc	Seeds
1.5	3	6	4	4	5	4	2	2	3
1.7	3	6	4	4	3	4	2	2	2
2	3	4	4	4	3	4	2	2	2
2.3	3	4	4	4	3	4	2	2	2
2.5	3	4	5	4	3	4	2	2	2

表 6　PE 在不同模糊加权指数下的最佳聚类数

m	DS1	DS2	DS3	DS4	DS5	DS6	Iris	Wdbc	Seeds
1.5	3	5	4	4	2	4	2	2	2
1.7	3	4	4	4	2	2	2	2	2
2	3	2	2	4	2	2	2	2	2
2.3	3	2	2	4	2	2	2	2	2
2.5	3	2	2	4	2	2	2	2	2

表 7　XB 在不同模糊加权指数下的最佳聚类数

m	DS1	DS2	DS3	DS4	DS5	DS6	Iris	Wdbc	Seeds
1.5	3	6	4	4	4	4	2	2	2
1.7	3	6	4	4	2	4	2	2	2
2	3	6	4	4	2	4	2	2	2
2.3	3	4	5	4	2	4	2	2	2
2.5	3	4	5	4	2	4	2	2	2

表 8　UV 在不同模糊加权指数下的最佳聚类数

m	DS1	DS2	DS3	DS4	DS5	DS6	Iris	Wdbc	Seeds
1.5	3	6	5	4	5	4	2	3	3
1.7	3	6	5	4	3	4	2	2	3
2	3	4	4	4	2	4	2	2	2
2.3	3	4	4	4	2	2	2	2	2
2.5	3	4	3	4	2	2	2	2	2

表 9　FM 在不同模糊加权指数下的最佳聚类数

m	DS1	DS2	DS3	DS4	DS5	DS6	Iris	Wdbc	Seeds
1.5	3	2	2	4	2	2	2	2	2
1.7	3	2	2	4	2	2	2	2	2
2	3	2	2	4	2	2	2	2	2
2.3	3	2	2	4	2	2	2	2	2
2.5	3	2	2	4	2	2	2	2	2

5　结束语

针对现有指标的缺陷，本文提出了 W 聚类有效性指标。根据在人工数据集和真实数据集上的实验结果表明， W 指标可以在有噪声和类间存在重叠的情况下作出正确判断，并且与模糊加权指数的相关性很小，性能稳定。由于 FCM 算法的处理对象主要针对团簇状分布数据，对非团簇状分布数据有效性差。在今后的研究过程中，可以针对非团簇状分布数据的有效性问题进行深入的研究。

参考文献：

[1] Zalik K R. Cluster validity index for estimation of fuzzy clusters of different [J]. Pattern Recognition, 2010, 43 (10): 3374-3390.

[2] Yang Lei. Extending information-theoretic validity indices for fuzzy clustering [J]. IEEE Trans on Fuzzy Systems, 2017, 25 (4): 1013-1018.

[3] Ruspini E H. A New Approach to Clustering [J]. InfCont, 1969, 15 (1): 22- 32.

[4] Dunn J C. A Fuzzy relative of the ISODATA process its use in detecting compact well-separated clusters[J].Journal of Cybernetics, 1974,3(3):32- 57.

[5] Le Capitaine H, Carl Freli. A cluster-validity index combiningan overlap measure and a separation measure based on fuzzy-aggregation operators [J]. IEEE Trans on Fuzzy Systems, 2011, 19 (3): 580-588.

[6] Bezdek J C, Life Fellow. The generalized c index forinternal fuzzy cluster validity [J]. IEEE Trans on Fuzzy Systems. 2016, 24 (6): 1500-1512.

[7] 高新波，谢维信. 模糊聚类理论发展及应用的研究进展 [J]. 科学通报, 1999, 44 (21): 2241-2251.

[8] 朴尚哲，超木日力格，于剑. 模糊 C 均值算法的聚类有效性评价 [J]. 模式识别与人工智能, 2015, 28 (5): 452-461.

[9] Tas demir K. A validity index for prototype-based clustering of data sets with complex cluster structures [J]. IEEE Trans on Systems, 2011, 41 (4): 1039-1053.

[10] 范九伦. 基于模糊熵的聚类有效性函数_范九伦 [J]. 模式识别与人工智能, 2001, 14 (4): 390-394.

[11] Liliane Silva. An Interval-based framework for fuzzy clustering applications [J]. IEEE Trans on Systems, 2015, 23 (6): 2174-2187.

[12] 毕凯. 基于模糊测度和证据理论的模糊聚类集成方法 [J]. 控制与决策, 2015, 30 (5): 823-830.

[13] Xie X L, Beni G. A validity measure for fuzzy clustering [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1991, 13 (8): 841-847.

[14] 张宇献，刘通. 基于改进划分系数的模糊聚类有效性函数 [J]. 沈阳工业大学学报, 2014, 36 (4): 431-435.

[15] 孟令奎，胡春春. 基于模糊划分测度的聚类有效性指标 [J]. 计算机工程, 2007, 33 (11): 15-17.

[16] Chen M Y, Linkens D A. Rule-base self-generation and simplification for data-driven fuzzy models [J]. Fuzzy Sets and Systems, 2004, 142 (2): 243- 265.

[17] Pal N R, Bezdek J C. On Cluster Validity for the fuzzy C-Means model [J]. IEEE Trans on Fuzzy Systems, 1995, 3 (3): 370-379.